Transcriptome Reconstruction from Single RNA-Seq Reads Using Expectation Maximization Algorithm with Expected Deviation Minimization Enhancement

Adrian Caciula[1], Serghei Mangul[2], Ion Mandoiu[3], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, GA 30303
Email: {acaciula, alexz}@cs.gsu.edu
[2] Department of Computer Science, University of California, Los Angeles, CA 90095
Email : serghei@cs.ucla.edu
[3] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269 Email : ion@engr.uconn.edu

High-throughput RNA sequencing (RNA-seq) [1], the latest technology for transcriptome analyses [2], allows to reduce the sequencing cost and significantly increase data throughput. However, as shown by recent studies [3], the data from RNA-seq is computationally challenging to use for reconstructing full-length transcripts due to sequencing errors and highly similar transcripts produced by alternative splicing. A number of recent works have addressed the problem of transcriptome reconstruction from RNA-Seq reads. In [4–6] the authors propose a "genome-guided" method that first map all reads to the reference genome using spliced alignment tools, such as TopHat [7], and then use the spliced reads to reconstruct the transcripts. The method of Trapnell et al. [4], referred to as Cufflinks, constructs a read overlap graph and generates candidate transcripts by finding a minimal size path cover via a reduction to maximum matching in a weighted bipartite graph.

In this paper we propose a statistical "genome-guided" method called "Expectation Maximization algorithm with **E**xpected **D**eviation **M**inimization Enhancement" (EM-EDM) for transcriptome reconstruction from single RNA-Seq reads. The first step of EM-EDM is to map the reads onto the genome using the spliced alignment tool, TopHat [7], as done, e.g., in [6, 4]. From mapped reads we first infer the exon-exon junctions to build a *splice graph* from which we enumerate all maximal paths corresponding to putative transcripts [8]. Next we compute the frequency (interchangeably referred in literature as expression levels) for all the candidate transcripts using EM algorithm. We further adjust the estimations by applying a novel EDM approach. The objective is to select the smallest set of putative transcripts that yields a good statistical fit between the exon-exon junctions and the expression levels of candidate transcripts.

The EM-EDM algorithm starts with the set of N known candidate transcripts and initialize their frequencies, f_t, with EM estimates. Similar transcript frequency estimations are computed by $IsoEM$ [9]. In addition, our method incorporates EDM, a fine tuning for frequency estimation which further improves

the accuracy of the computation. Reducing the error rate is critical for detecting similar transcripts especially in those cases when one is a subeset of another.

Let l_t be the adjusted length of the transcripts $t \in H$ (i.e., the length of t minus the average fragment length), where H is the set of all candidate transcripts. The expected read frequency e'_i equals

$$\begin{aligned} s_r &= \sum_{t \in H} f_t h_{t,r} \\ q_t &= \sum_{r \in R} \frac{h_{t,r}}{s_r} \\ e'_r &= \frac{1}{s_r} \sum_{t \in H} \frac{l_t f_t h_{t,r}}{q_t} \\ e_r &= \frac{e'_r}{\sum_{r \in R} e'_r} \end{aligned} \tag{1}$$

o_r for each read was computed as in [9].

The transcript frequency can be estimated by the following iterative process. Given transcript frequencies estimates, signed deviation of expected from observed read frequencies, the algorithm increments and decrements transcript frequencies in order to decrease the total deviation.

Expected Deviation Minimization method (EDM). Initialize $f_t \leftarrow$ corresponding EM frequency Set $D \leftarrow 1$ and $C \leftarrow 0.05$. Each iteration consists of the following three steps:

- Estimate the expected read frequency e_r for each read r according to (1) and its deviation $d_r = e_r - o_r$, $D^{next} = \sum_{r \in R} |d_r|$. If $D^{next} > D$, then $C \leftarrow C/2$ and recompute e_r and d_r else update $f_t \leftarrow f_t^{next}$, $D \leftarrow D^{next}$.
- Estimate signed deviation from expected haplotype frequency for each transcript t:

$$D_t^0 = \frac{1}{l_t} \sum_{r \in R} h_{r,t} d_r \tag{2}$$

$$D_t^1 = D_t^0 - \frac{\sum_{t \in T} D_t^0}{|T|} \tag{3}$$

$$D_t^2 = C \times \frac{D_t^1}{\sum_{t \in T} |D_t^1|} \tag{4}$$

- Update transcript frequency estimation

$$f_t^{next} = f_t - D_t^2 \tag{5}$$

Iterations are repeated until $C < \epsilon_a$. Let $\epsilon_a = 0.005$

The formula (2) finds scaled deviation of each t, (3) centralizes deviation and (4) says that the total update does not exceed C.

To filter the set of candidate transcripts, we estimate the frequency for each of them using our EM-EDM approach, then based on their frequency priority we select one by one until all exon-exon junctions found during mapping are covered.

For our simulations we have used Human genome $UCSC$ annotations, GNFAtlas2 gene expression levels with uniform expression of gene transcripts. We have simulated uniform single-end reads of length $100bp$, coverage $100x$, and a fragment length distribution of $500bp$ with a standard deviation of $50bp$.

Following [10], we use sensitivity and Positive Predictive Value (PPV) to evaluate the performance of different methods. Sensitivity is defined as portion of the annotated transcript sequences being captured by candidate transcript sequences as follows:

$$Sens = \frac{TP}{TP + FN}$$

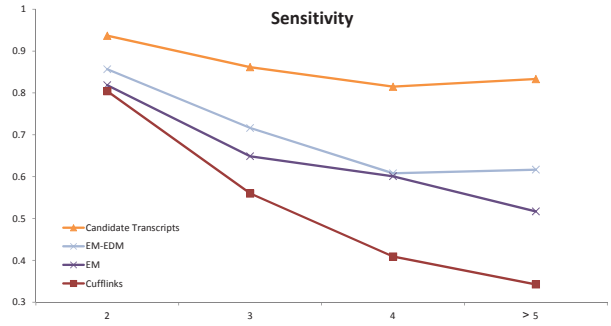
PPV is defined portion of annotated transcript sequences among candidate sequences as follows:

$$PPV = \frac{TP}{TP + FP}$$

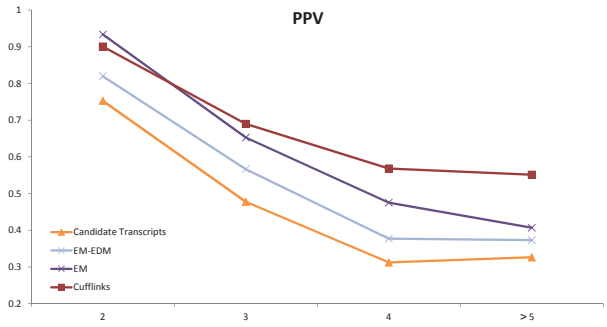
Figure 1 shows our preliminary experimental results on synthetic datasets where we can observe that EM-EDM has increased transcriptome reconstruction sensitivity compared with Cufflinks. "Candidate Transcripts" represent all maximal paths corresponding to putative transcripts built from our *splice graph*. For future work we plan the improve the filtering algorithm in order to increase the PPV.

References

1. A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods*, 2008. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1226
2. Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics." *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009. [Online]. Available: http://dx.doi.org/10.1038/nrg2484
3. T. R. Mercer, D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick, and J. L. Rinn, "Targeted RNA sequencing reveals the deep complexity of the human transcriptome." *Nature Biotechnology*, vol. 30, no. 1, pp. 99–104, 2012.
4. C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010. [Online]. Available: http://dx.doi.org/10.1038/nbt.1621
5. W. Li, J. Feng, and T. Jiang, "IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly," *Lecture Notes in Computer Science*, vol. 6577, pp. 168–+, 2011.



(a)



(b)

Fig. 1. Flowchart for ML-EDM: (a) Sensitivity and (b) Positive Predictive Value(PPV)

6. M. Guttman, M. Garber, J. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. Koziol, A. Gnirke, C. Nusbaum, J. Rinn, E. Lander, and A. Regev, "*Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, no. 5, pp. 503–510, 2010. [Online]. Available: http://dx.doi.org/10.1038/nbt.1633
7. C. Trapnell, L. Pachter, and S. Salzberg, "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp120
8. S. Mangul, A. Caciula, S. A. Seesi, D. Brinza, A. R. Banday, R. Kanadia, I. Mandoiu, and A. Zelikovsky, "Flexible approach for novel transcript reconstruction from rna-seq data using maximum likelihood integer programming," in *Proc. 5th International Conference on Bioinformatics and Computational Biology*, March 4-6, 2013 2013, p. to appear. [Online]. Available: http://bicob.ece.iastate.edu/
9. M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from rna-seq data," *Algorithms for Molecular Biology*, vol. 6:9, 2011. [Online]. Available: http://www.almob.org/content/6/1/9
10. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC Bioinformatics*, vol. 12, no. Suppl 6, p. S1, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S6/S1